The National Digital Newspaper Program (NDNP)
Technical Guidelines for Applicants

Overview of Technical Approach for Phase I

The National Digital Newspaper Program is a long-term effort and the technical environment will change as the program continues. The National Endowment for the Humanities (NEH) and the Library of Congress (LC) have selected a technical approach for Phase I to balance long-term objectives and shorter-term constraints. These include:

- convenient accessibility over the World Wide Web for the general public to the entire collection as it grows, through a consistent interface and using proven technology;
- page images of sufficient spatial and tonal resolution to support effective performance of OCR (optical character recognition) software and representation of printed half-tones, given the limitations of microfilm, expecting that future improvements in OCR and image processing will be applied to the same images;
- the use of digital formats with a high probability of sustainability - in particular, using standard formats where possible and proprietary formats only where widely adopted;
- and attention to the cost of digital conversion and maintenance of the resulting assets.

The goal of the initial program phase is to build a Web-accessible NDNP delivery application with sufficient geographic coverage and digital assets to validate the technical approach and to serve as a test bed for future research and development for techniques to enhance the content and access interface, to support effective use by scholars and the general public.

In succeeding phases of the project, the approach and associated guidelines will be evaluated and revised based on feedback from awardees, experience in providing access to historic newspapers online, and technological advances.

In summary, the approach for Phase I is based on:

- grayscale images (generally 400 dpi) from microfilm for newly-converted materials (see Scanning section for details on re-purposing existing digital collections)

- OCR with character or word-bounding boxes, uncorrected, with recognition of columns, but without segmentation of pages into articles,
- structural metadata for pages, issues, editions, and titles to support a calendar-based browsing interface,
- copies of all page images and associated metadata at LC,
- an interface designed specifically for access to historic newspapers in the public domain, mounted at LC (the initial interface will permit full-text searches with retrieval of individual page images, and highlighting of search words on the images), and
- the ability of awardees to re-use any digital assets created for NDNP in other systems or for other purposes.

NEH and LC recognize that other institutions may choose other approaches or formats for their own digital repository and delivery systems and thus either weigh costs and benefits differently or wish for compatibility to existing systems. Applicants are encouraged to pursue local approaches in parallel with participation in NDNP, with the overall goal of providing effective widespread access to newspapers through scanning and text conversion and evaluating alternative interfaces for navigating and exploring large collections of newspapers. Applicants who use other formats locally should be capable of providing digital assets to the NDNP according to the specifications described below.

The National Digital Newspaper Program supports a consistent technical specification for digital newspaper reproductions and associated metadata in order to maintain parity of services for materials from a variety of institutions and collections and to support the "best practices" of today's understanding of digital preservation needs.

Deliverables

Awardees are expected to deliver the following to the Library of Congress, to allow construction of a permanent archive and a unified interface for searching and browsing the entire NDNP collection. After the cooperative agreements are announced, LC will convene a meeting of awardees to review these technical guidelines, and establish work-plan milestones, and specifications for deliverables.

For each state

- Essay – history of state's newspapers for the relevant time period (1900-10) – 1,000 words.

For each title

- Up-to-date MARC record from the CONSER database, fully conformant to current standards for cataloging print newspapers,
- Additional title-level metadata related to the title run/s digitized and delivered (see Appendix A: Digital Asset Metadata Elements), and
- Essay – scope and content of each title, history and significance – 250 words.

For each issue/edition

- Structural metadata for issues/editions digitized and organized by date (see Appendix A: Digital Asset Metadata Elements)

For each newspaper page

- Page image in two raster formats
  - Grayscale, generally 400 dpi, uncompressed TIFF 6.0 (see Scanning below),
  - Same image, compressed as JPEG2000 (see Scanning below),
- File with OCR text and associated bounding boxes for words or characters (see OCR details below), 1 file per page image,
- PDF Image with Hidden Text, i.e., with text and image correlated (see OCR details below),
- Structural metadata to relate pages to title, date, and edition, sequence pages within issue or section; and to identify image and OCR files (see Appendix A: Digital Asset Metadata Elements), and
- Technical metadata to support the functions of a trusted repository (see Appendix A: Digital Asset Metadata Elements and Appendix C – TIFF 6.0 Header Tags).

During Phase I of the program, awardees will have flexibility as to the form in which metadata is delivered (XML, tab-delimited files, MySQL, or MS Access databases, etc.). This phase will be an opportunity to establish a standard form for the metadata based on practical experience.

The awardee shall organize the page images and related files for each newspaper title in a hierarchical directory structure sufficient for identification of the individual digital assets from the metadata provided.

Options and specifications for delivery will be specified at the initial awardees' meeting, post-award.

For each microfilm reel digitized:

- A second-generation (2N) duplicate silver negative microfilm, made from the camera master, will be barcoded and deposited with the Library of Congress on completion of the award (LC to supply barcodes for all reels), and
- Technical metadata concerning the quality characteristics of the film used for digitization (See Appendix B – Microfilm Quality Metadata).

Supporting documentation

OCR output: Since no standard format exists for OCR output and the representation of associated bounding-box data, awardees should provide LC with documentation showing how the OCR output elements are represented in the digital files delivered. This requirement may be met by providing appropriate documentation for the OCR software used.

Metadata syntax and files: Awardees will provide detailed documentation about the form and structure in which metadata is delivered (e.g., field names, file format, character encoding, etc.).

Technical Details

Selection

The goals of the overall project, the chronological scope (1900-10), and the intellectual criteria for selecting newspaper titles for Phase I are described in the Request for Proposal guidelines. To ensure the highest quality and most usable digital products and services, the process for selection of a newspaper title for inclusion in the NDNP should also incorporate a technical analysis of the microfilm to be scanned.

For NDNP (and the associated collection of duplicate microfilm negatives (2N)) to be as complete as possible, the following guidelines should be followed:

1. Complete (or majority of) title run should be available on microfilm without restrictions that interfere with the goals of the program;

2. An effort should be made to deliver as complete a title run as possible. Locating and substituting a limited number of scanned images from paper may be necessary to complete the run.

Several technical factors will affect the success of microfilm scanning and optical character recognition (OCR). The following factors should be considered during the selection process. They include:

1. The quality of original text and microfilm capture. Poorly prepared original material, no matter how well microfilmed, yields poor results. Microfilm of bound material may have page curvature, gutter shadows, or out of focus pages that influence digital image quality. Preference in selection should be given to titles on higher quality microfilm.

2. The reduction ratio used when microfilming the original newspaper. This ratio directly influences image quality and OCR results. The lower the reduction ratio (below 20x) the better. (If the reduction ratio is too high to allow scanning at 400 dpi, tests on sample images should be performed to determine if a lower resolution (e.g., 300 dpi) provides acceptable confidence levels in OCR text.)

3. The camera master negative microfilm duplicated for scanning should have resolution test patterns readable at 5.0 or higher. For camera master microfilm without resolution test charts, resolution should be estimated by comparison to film with resolution test charts and original material.

4. Variations in density within images and between exposures. Such variations require adjustment of scanning parameters within a reel. Density readings should follow current standards, but the range should ideally be narrower than the standards allow (e.g. .90-1.20). Best results are obtained from microfilm with variations in density readings of no more than 0.2 within an image and between exposures.

5. Confidence level through OCR testing of sample page images. Searchable text using OCR is a key element of NDNP. For a camera master negative that is questionable with respect to any of the above criteria (resolution, reduction ratio, densities, etc.), sample digital images may need to be tested for acceptable OCR confidence levels to determine suitability for selection.

Note: The current guidelines for microfilming newspaper for the USNP are available at http://www.loc.gov/preserv/usnpspecs.html.

Scanning and Master Image Format

For Phase I and based on preliminary analysis, scanning specifications should follow these guidelines:

- scan from a clean second-generation duplicate silver negative microfilm (to be deposited with the Library of Congress at the end of the award period);
- capture specifications are 8-bit grayscale at a resolution of 400 dpi, if possible (relative to the physical dimensions of the original newspaper, rather than the microfilm). For the scanner operator to achieve this, the microfilm reduction ratio must be known or derived by other means;
- a standards-based target film strip should be scanned at the start of each session, to monitor scanning equipment performance. Target test images should be delivered along with the page images;
- provide the master page images, delivered to LC, as uncompressed images in TIFF 6.0 format.

NOTE: Existing digital page images may be re-purposed for delivery to NDNP (up to 20,000 pages). Consideration should be based on the acceptability and accuracy of OCR produced from these existing images (propose re-scanning to current specifications, if necessary to accomplish acceptable results.) Except for scanning specifications, these re-purposed materials must meet all other technical guidelines described below, producing the same deliverables. Do not modify imaging specifications on existing assets (e.g. scale spatial resolution to 400 dpi) to meet NDNP specifications.

Newspapers microfilmed two sheets per frame should be split into two separate image files (and assigned appropriate metadata). As grayscale images are requested, despeckling, edge-enhancement, and similar techniques that are relevant when producing bitonal images should not be necessary. To improve appearance and OCR accuracy, images with more than 3 degrees of skew should be deskewed. Image files should be cropped to the page edge (not to the text block boundaries). All operations that change

the image dimensions, spatial resolution, or orientation (e.g., cropping, deskewing) must be made before OCR, since the OCR output is expected to include bounding-box coordinates to relate words and characters to their position on the page in the search interface. If these enhancements are integrated into the OCR operation as a pre-processing step, the grayscale image files delivered to LC should have the same enhancements.

Microfilming target frames should be captured as images and delivered with other digital assets. Such images will be treated as digital assets for archiving but not normally displayed in the NDNP access interface, as they represent an artifact of the microfilming process rather than intellectual content of the collection. If convenient, these can be delivered as bitonal TIFFs with Group 4 compression.

In addition, a standards-based scanning target film strip should be scanned at the start of each session, to monitor scanning equipment performance. Target test images should be delivered along with the page images. Specific test targets and quality analysis tools will be discussed with awardees at the post-award awardees' meeting.

For the NDNP access interface, LC has developed and expects to employ a zooming capability based on JPEG2000 wavelet compression. This technology will not only compress the newspaper image effectively but also permit the presentation of image segments dynamically, at the user's request.

LC and NEH intend to implement the draft standard NISO Z39.87 –2002 Data Dictionary – Technical Metadata for Digital Still Images for master images in the NDNP. To support LC's responsible custodianship of these images, the headers for the master TIFF images should incorporate tagged metadata relating to the creation of the images.

Awardees will incorporate:

- tags usually required of LC contractors (see http://memory.loc.gov/ammem/prpsal/attach5.html),
- tags corresponding to mandatory elements from the draft standard NISO Z39.87 – 2002, if the elements are both relevant for these uncompressed grayscale images and appropriate TIFF, TIFF/EP, or EXIF tags exist, and
- additional tags corresponding to recommended elements from Z39.87 –2002 and identified as significant for NDNP.

Appendix C lists appropriate tags identified by the Library of Congress. Additional tags necessary for rendering of the image (e.g., tile specifications, if used) should be present, as well. Image headers may contain additional compatible tags.

LC and NEH recognize that NISO Z39.87 is a Draft Standard and that scanning vendors may not be equipped to provide all its metadata elements. Awardees should nevertheless use their best efforts to have this metadata correctly embedded in the TIFF headers.

Summary of Scanning Guidelines

1. Digital reproductions should be made from a preservation copy of microfilm, a clean second-generation duplicate silver negative.

2. Technical recommendation: 400 dpi (relative to original materials), 8-bit grayscale, TIFF 6.0 uncompressed. (Re-purposed digital images from existing projects may vary.)

3. Two-up film should be split so that there is one page image per file.

4. De-skew images with a skew of greater than 3 degrees. (Greater skew leads to less accurate OCR.)

5. Crop to edge of page (optional for re-purposed materials)**.**

6. Capture microfilm target frames (optional for re-purposed materials). These image files to be identified as "targets" in metadata; will not be used for display.

7. Capture scanning resolution targets at the start of each session, to monitor scan quality (optional for re-purposed materials). These targets should be delivered with microfilm targets and page images.

Note: the grayscale images sent to LC must have exactly the same dimensions, spatial resolution, skew, and cropping as the images used for OCR.

OCR and Associated Information

Machine-readable text allows users to search a newspaper or a collection of newspapers for names of people and places, and for phrases, and provides the potential to use more powerful data-mining or natural language analysis techniques to locate relevant articles. The provision of machine-readable text correlated with page images is a tremendous aid to users seeking to navigate the complicated layouts and large, text-intensive pages of newspapers. It permits the examination of the relationships between various articles, visually and textually. Development of the Phase I NDNP access interface will be based on a fully automated approach to text conversion without article-level segmentation or article-level metadata.

OCR creates machine-readable text from scanned page images and permits full-text searching of the contents of newspaper pages. Bounding-box data relates characters or words recognized to their position on the image. Coordinates describe the position and outer dimensions of a box enclosing a character or word, and/or space(s), in the original image. The initial NDNP application will search uncorrected OCR text at the page-level, using bounding-box coordinates for characters (preferably) or words to correlate text elements to position on the page, so that search words can be highlighted in the interface.

Bounding-box data typically includes four numbers. One approach provides the x^1 x^2 and y^1 y^2 coordinates, measuring the horizontal (*x*) and vertical (*y*) distance of the top left and the bottom right corners of the box, in points, pixels, or inches, from the top left corner of the image. An alternative and equivalent approach provides *x*,*y*,*w*,*h*, with the *x*,*y* coordinates representing the horizontal and vertical distance from the top left of the image to the top left corner of the bounding box, and *w* and *h* representing the width and height of the box.

Important: The page images delivered must correspond in dimensions, orientation, and skew to those used for the OCR. **Once a page image is processed by the OCR application, the resulting text must not be changed**. Text corrections will dissociate the required bounding-box layout coordinates from the appropriate words or characters, resulting in incorrect representation (e.g., misaligned highlighting) in the NDNP access interface.

Summary of OCR Guidelines

Required elements for OCR files:

1. One OCR text file per page image. (Discrete files should be produced for each page, rather than for a multi-page issue or entire title).
2. Each OCR text file name corresponds to the page image it represents.
3. Text in ASCII or UTF-8 character set.
4. No graphic elements saved with the OCR text.
5. OCR text ordered column-by-column (that is, in a natural reading order).
6. OCR text file with bounding-box coordinate data at the character or word level.
7. OCR software documentation for the representation of all elements listed here (e.g., the bounding-box coordinates) provided, showing how the elements are represented in the actual digital files.

If possible, additional elements for OCR files:

1. Confidence level data at the page, line, character, and/or word level.
2. Point size and font data at the character or word level.

Note: Zones for articles will not be used in the initial interface, but if the OCR process selected by an awardee does generate coordinates for zones, the feasibility of supplying this information in a form convenient to an awardee and to LC will be explored during the planning stage of the project.

Other Derivative Files

In addition to the text file and bounding-box coordinate data, the awardee institution will provide a searchable PDF (Portable Document Format) Image with Hidden Text for each page image and a JPEG2000 compressed image file (.JP2).

PDFs will provide an image of the original page that can be conveniently printed and support within-page searching for words. LC will use the separate OCR output file as the

basis for search in its access interface. The PDF Image with Hidden Text can be created at the time of processing by the OCR application.

Required elements for PDF files

1. PDF Image with Hidden Text for each page image.
2. Each searchable PDF file name corresponds to the page image it represents.
3. The PDF files should incorporate appropriate metadata to allow users of downloaded images to identify the source publication, date and page number.

The objective of requesting JPEG2000 compressed image files is to provide a high-quality, low-bandwidth presentation of the page image that can be stored independently of the large master digital files. Use of JPEG2000, Part 1, (or ISO-15444) is planned for the initial NDNP interface. The .JP2 compression options and guidelines, along with other technical specifications, will be discussed at the post-award awardees' meeting.

Metadata

One aim of the LC/NEH partnership in establishing the National Digital Newspaper Program is to integrate historical newspaper collections digitized by many institutions into a single searchable resource, allowing users to search across multiple titles with a single query. To achieve this while allowing institutions the flexibility to incorporate materials into their own catalog systems and online services, NDNP awardees must deliver to LC copies of descriptive records from CONSER and metadata for various levels of granularity within the digital reproductions.

Each newspaper digitized through NDNP must be supported by coherent metadata, to provide intellectual access and support navigation of the structure of the publication, by date, section, etc. The tables in Appendix A list the elements appropriate at the newspaper title level, the issue/edition level, and the page level. [The tables indicate whether elements are mandatory or repeatable.] The access interface will permit direct identification and citation at each level through persistent identifiers. The identification of newspapers titles will be based on LCCNs, since not all historical newspapers have ISSNs. The final metadata specifications will be discussed at the post-awardees' meeting.

During Phase I of the program, awardees have flexibility as to the form in which metadata is delivered (XML, tab-delimited files, MySQL, or MS Access databases, etc.). This phase provides an opportunity to establish a standard form for the metadata based on practical experience.

All newspaper titles in NDNP must be described in CONSER records, with a full bibliographic record at the title-level for the original materials. If pre-existing, the CONSER records must be reviewed and updated as necessary by the awardee institution and copies delivered with the project data. The records should be in MARC 21 Communications format or converted to MARC XML with UTF-8 character encoding.

Summary of All Digital Asset Deliverables

1. Master digital page image format = TIFF 6.0 uncompressed,
2. OCR text file with bounding-box coordinates = 1 text file per page,
3. PDF Image with Hidden Text = 1 PDF per page, and
4. Derivative digital page image format = JPEG2000 (.JP2) using specified compression options,
5. Metadata in accordance with guidelines in Appendix A (all records should be combined into one dataset).

Note: The four digital files associated directly with a newspaper page (.TIF, .JP2, .PDF, and OCR) are expected to use the same file identifiers with distinct file extensions.

Appendices

Legend for tables:

M	Mandatory	R	Repeatable
MA	Mandatory if applicable	NR	Non-repeatable
O	Optional		

Note: shaded rows (light blue) are elements to be created by LC at the time of ingest.

Appendix A: Digital Asset Metadata Elements

TITLE DATA

For each title

Title will correspond to a single MARC record in CONSER (itself related to holdings information in the USNP UnionList, maintained by OCLC). Most metadata elements at title level will be drawn from the CONSER record. Awardees are responsible for ensuring the quality of the CONSER records for the titles selected for digitization and for delivering copies to the Library of Congress if copies of the record are not found in the Library of Congress OPAC.

Element	Data Type	Example	Notes	From CONSER record	Repeatable	Mandatory
LCCN	string	sn83031150	LCCN used as the basis for a title identifier instead of ISSN because not all newspapers have ISSNs. MARC 010. Use canonical form of LCCN. See http://www.loc.gov/marc/lccn-namespace.html.		NR	M
OCLC Number	string		MARC 035 (for matching to CONSER record).	X	NR	M
Title	string	Brooklyn eagle	MARC 130, if present. Otherwise use MARC 245 $a.	X	NR	M
ISSN	string	0145-0808	MARC 022 $a.	X		MA
Repository for Source Microfilm	string	Photoduplication Services, Library of Congress, Washington, DC 20540	Corresponds to information in MARC 852 $a, b, and e, with added punctuation.		NR	M
Repository Code	string	DLC	From MARC Organization Code list.		NR	M

Element	Data Type	Example	Notes	From CONSER record	Repeatable	Mandatory
Responsible Institution (Digital)	string	Library of Congress, Washington, DC 20540	Institution providing digitization services.		NR	M
Responsible Institution Code	string	DLC	From MARC Organization Code list.		NR	M
Start Date (for digitized run)	date	1903-02-04	International Standard ISO 8601. http://www.w3.org/TR/xmlschema-2/#date. The lexical representation for date is the reduced (right truncated) lexical representation for dateTime: CCYY-MM-DD. [A single lexical representation, which is a subset of the lexical representations allowed by [ISO 8601], is allowed for dateTime. This lexical representation is the [ISO 8601] extended format CCYY-MM-DDThh:mm:ss where "CC" represents the century, "YY" the year, "MM" the month and "DD" the day, preceded by an optional leading "-" sign to indicate a negative number.]		NR	M
End Date (for digitized run)	date	1905-03-15	International Standard ISO 8601. http://www.w3.org/TR/xmlschema-2/#date.		NR	M

Element	Data Type	Example	Notes	From CONSER record	Repeatable	Mandatory
Issue/Edition Pattern	string	Daily (except Sunday) with regional editions on Sat.	Should reflect pattern digitized and delivered. Usage in MARC 310, 321 can serve as guide.			M
Title Identifier	anyURI	http://hdl.loc.gov/loc.ndnp/sn83031150	Title level identifier for citation – to be assigned by LC. *[The example should not be interpreted as the pattern that will be used.]*			LC

Subsidiary Information for Title: Geographic Coverage

For current administrative and geographic boundaries, supply this information for each county or significant city for which the newspaper provided local coverage.

Element	Data Type	Example	Notes	Repeatable	Mandatory
LCCN	string	sn83031150	To relate to Title.	NR	M
Coverage Country	string	United States	Use values as for MARC 752 $a.		M
Coverage State	string	New York	Use values as for MARC 752 $b.		M
Coverage County	string	Kings	Use values as for MARC 752 $c.		O
Coverage City	string	Brooklyn	Use values as for MARC 752 $d.		O

ISSUE/EDITION DATA

For each issue/edition expected in the publication pattern, including missing issues/editions.

{LCCN, Date, Edition Order} will form identifier for edition (to which pages must relate). LC will supply barcodes for all duplicate microfilm reels. Provide issue/edition records for all **known** issue/edition occurrences (i.e. if microfilm reel includes information

(target or Guide to Contents) indicating an issue/edition was known to be published but is not available as a digital asset at this time, create a record for that issue/edition and use the Issue Present Indicator to indicate the issue/edition the record describes is not available.)

Field Name	Data Type	Example	Notes	Repeatable	Mandatory
LCCN		sn83031150	To relate to Title.	NR	M
Reel Number	string	375892205698	Reel number corresponds with LC barcode supplied for all duplicate microfilm reels deposited with LC. *Not mandatory for re-purposed materials.	NR	M**A**
Issue Present Indicator	boolean	1	Valid numeric values are: 1 - Issue published and digitized pages present. 0 - Issue not digitized.	NR	M
Issue Present Comment	string	No issue published due to weather.	To record any additional known information indicated in film on missing issues. Default value if not empty = "Filmed target indicates issue missing." Include more specific information, if known.	NR	MA
Date	date	1908-01-01	Actual date issued, corrected if necessary. ISO 8601 style: CCYY-MM-DD.	NR	M
[Incorrect] Date as Labeled	date	1907-01-01	If date printed was in error (not the date issued), this field reflects the incorrect date as printed. Otherwise leave blank.	NR	MA

Field Name	Data Type	Example	Notes	Repeatable	Mandatory
Volume Number	string	27	Following SICI standard: 1. All numeric information shall be converted to arabic numbers. 2. Alphabetic data used as enumeration designations shall be transcribed as they appear on the piece, and converted to uppercase.	NR	O
Issue Number	string	3	Following SICI standard: 1. All numeric information converted to arabic numbers. 2. Alphabetic data used as enumeration designations transcribed as they appear on the piece and converted to uppercase.	NR	O
Edition Label	string	Late City Final	If present, as printed.	R	MA
Edition Order	positive integer	1	Default is 1. If more than one edition on this date and LCCN is known, number in chronological order.	R	M
Issue Identifier	anyURI	http://hdl.loc.gov/loc.ndnp/sn83031150.19080101	Issue level identifier for citation – assigned by LC. *[The example should not be interpreted as the pattern that will be used.]*		LC

PAGE DATA

For each newspaper page expected in the publication pattern, including missing pages and filmed targets.

Provide page records for all **known** page occurrences (i.e. if microfilm reel includes information (target or Guide to Contents) indicating a page was known to be published but is not available as a digital asset at this time, create a record for that page and use the Page Present Indicator to indicate the page the record describes is not available.)

Field	Data Type	Example	Notes	Rep	Mand
LCCN	string	sn83031150	To relate to Title, see Title table.	NR	M
Date	date	19080101	To relate to Issue Date, see Issue/Edition table.	NR	M
Edition Order	positive integer	1	To relate to Edition Order, see Issue/Edition table.		
Reel Number	string	375892205698	Reel number to correspond with LC barcode supplied for all duplicate microfilm reels deposited with LC. *Not mandatory for re-purposed materials.	NR	M**A**
Page Present Indicator	enumeration	1	This field is to indicate a page that was not available to microfilm. Valid numeric values are: 1 - Page image present. 2 - Page image known to be missing. 9 – Image is filmed target.	NR	M
Record Sequence Number	positive integer	13	This orders the records for page records within an issue – useful for multi-section issues.	NR	M
Section Label	string	B	If present, as printed. Could be blank, “C,” “IV,” “3,” “Business,” etc.	NR	MA
Page Number	string	B3	Exactly as it appears on the page.	NR	MA
File Name on Delivery	string	\sn83031150\1908\00000013	Derivatives and associated files should use same file name prefix as the TIFF. Extensions for TIFF, JPEG2000, and PDF files are assumed to be .TIF, .JP2, .PDF. Extension used for	NR	M

Field	Data Type	Example	Notes	Rep	Mand
Media			OCR output file should be supplied as part of supporting documentation. *[The example should not be interpreted as the pattern that will be used.]*		
Page Identifier	anyURI	http://hdl.loc.gov/loc.ndnp/sn83031150.1908010101013	Page level identifier for citation –assigned by LC. *[The example should not be interpreted as the pattern that will be used.]*	NR	LC

Appendix B: Microfilm Reel Quality Analysis Metadata Elements (for Newly-Converted Materials)

This information will be utilized to study the technical characteristics of microfilm used for digitization and they impact OCR quality, and usability, using a quantifiable approach.

Element	Data Type	Example	Notes
Title/s	string	National Forum	MARC 130, if present. Otherwise use MARC 245 $a. List multiple titles, as necessary, delimited by semi-colon. (;)
Reel Number	positive integer	375892205698	Reel number corresponds with LC barcode supplied for all duplicate microfilm reels deposited with LC.
Start Date	date	1910-05-28	Date/Time
End Date	date	1910-11-12	Date/Time
Position	string	2a	1a, 2a, 1b, 2b
Reduction Ratio	string	20x	If stated
Capture Resolution_Original	string	300dpi	Resolution relative to original material, measured in pixels/inch (or mm.).
Capture Resolution_Film	string	6000dpi	Resolution relative to microfilm, measured in pixels/inch (or mm.). Capture Resolution_Film = Reduction Ratio x Capture Resolution_Original.
Guide to Contents Present	boolean	1	Valid values are 1 (present), 0 (missing).
Guide to Contents String	string	Title established April 30…	If present, transcribe text from Guide to Contents, as it appears on film
Date Created	string	1986	Date of microfilm creation
Loose Leaves	boolean	0	For original material; Valid values are 1 (yes), 0 (no).
Bound Volume	boolean	1	For original material; Valid values are 1 (yes), 0 (no).
Comments	string		

Element	Data Type	Example	Notes
Dimensions	string	17x23 in.	From original materials, if possible, or from N.W. Ayer & Son. *N.W. Ayer & Son's American Newspaper Annual*. Philadelphia : N.W. Ayer and Son, 1880-1909.
Pages per Issue	positive integer	4	Estimate from microfilm, or from N.W. Ayer & Son. *N.W. Ayer & Son's American Newspaper Annual*. Philadelphia : N.W. Ayer and Son, 1880-1909.
Number of Resolution Targets	positive integer	5	
Resolution of Master	floating point number	7.1	
Resolution Comment Master	string		
Density Reading 1 Master	floating point number	0.91	
Density Reading 2 Master	floating point number	0.94	
Density Reading 3 Master	floating point number	1.01	
Density Reading 4 Master	floating point number	0.91	
Density Reading 5 Master	floating point number	0.95	
Density Reading 6 Master	floating point number	0.88	
Density Reading 7 Master	floating point number	0.92	
Density Reading 8 Master	floating point	1.01	

Element	Data Type	Example	Notes
	number		
Density Reading 9 Master	floating point number	0.95	
Density Reading 10 Master	floating point number	1.04	
Average Density Master	floating point number	0.95	
Dmin Master	floating point number	0.20	
Resolution of Duplicate Negative	floating point number	6.3	Film resolution of preservation copy of microfilm, a clean second-generation duplicate silver negative used for digitization.
Resolution Comment Duplicate Negative	string	Weak 7.1	
Density Reading 1 Duplicate Negative	floating point number	1.14	
Density Reading 2 Duplicate Negative	floating point number	1.17	
Density Reading 3 Duplicate Negative	floating point number	1.01	
Density Reading 4 Duplicate Negative	floating point number	1.15	
Density Reading 5 Duplicate Negative	floating point number	1.06	
Density Reading 6 Duplicate Negative	floating point number	1.13	
Density Reading 7 Duplicate Negative	floating point number	1.08	
Density Reading 8	floating point	1.02	

Element	Data Type	Example	Notes
Duplicate Negative	number		
Density Reading 9 Duplicate Negative	floating point number	1.18	
Density Reading 10 Duplicate Negative	floating point number	1.12	
Average Density Duplicate Negative	floating point number	1.10	
Dmin Duplicate Negative	floating point number	0.15	

Appendix C: TIFF 6.0 Header Tags

To ensure that LC the long-term sustainability and custody of master images created, the NDNP proposes to use Draft NISO Z39.87 – 2002 Standard, *Data Dictionary—Technical Metadata for Digital Still Images* as guidance. For Phase I, the elements identified in the table below will be incorporated into headers for the master TIFF images.

These elements include:

- tags routinely required by LC (see http://memory.loc.gov/ammem/prpsal/attach5.html),
- tags corresponding to elements from the Draft NISO Z39.87 –2002 Standard considered mandatory for uncompressed grayscale images and for which appropriate TIFF, TIFF/EP, or EXIF tags exist, and
- additional tags corresponding to recommended elements from Z39.87 –2002 and determined significant to NDNP for identifying page-images and recording key details about their creation.

Any other tags essential for rendering the uncompressed grayscale images should also be present (e.g., tile specifications, if tiles are used). Where no guidance is provided in the table, "typical" or "expected" values should be provided.

LC anticipates reviewing these specifications with awardees at a meeting of Phase I participants after awards are made.

Z39.87 #	Z39.87 Name	TIFF Tag #	TIFF Tag Name	Value	Notes
		269	DocumentName		Microfilm reel # (barcode).
6.1.3.1	CompressionScheme	259	Compression	1	
6.1.4.1	ColorSpace	262	PhotometricInterpretation	0 or 1	
6.1.5.2	StripOffsets	273	StripOffsets		
6.1.5.3	RowsPerStrip	278	RowsPerStrip		
6.1.5.4	StripByteCounts	279	StripByteCounts		
6.2.1	ImageIdentifier	42016	UniqueImageID		Image ID as file name or path. Must be unique within reel.
6.2.4	Orientation	274	Orientation		
7.1	SourceType	41728	FileSource	microfilm	Use other appropriate value if replacement image is scanned from other

Z39.87 #	Z39.87 Name	TIFF Tag #	TIFF Tag Name	Value	Notes
					medium.
7.3	ImageProducer	315	Artist		Institution name followed (if applicable) by ";" and name of scanning contractor.
7.6.1.1	ScannerManufacturer	271	Make		
7.6.1.2.1	ScannerModelName	272	Model		Include model number.
7.6.2.1	ScanningSoftware	305	Software		Include version.
7.9	DateTimeCreated	306	DateTime		
8.1.2	SamplingFrequencyUnit	296	ResolutionUnit	1 = no absolute unit of measurement 2 = inch	Value 2 indicates 282 and 283 are explictly expressed in pixels per inch (ppi).
8.1.3	XSamplingFrequency	282	XResolution		Resolution relative to size of original newspaper in pixels per inch. Approximation is acceptable.
8.1.4	YSamplingFrequency	283	YResolution		Resolution relative to size of original newspaper in pixels per inch. Approximation is acceptable.
8.1.5	ImageWidth	256	ImageWidth		In pixels.
8.1.6	ImageLength	257	ImageLength		In pixels.
8.2.1	BitsPerSample	258	BitsPerSample	8	
8.2.2	SamplesPerPixel	277	SamplesPerPixel	1	

Note: In addition to spatial resolution relative to the original (see TIFF tags 282 and 283) awardees will provide the resolution relative to the microfilm scanned as part of the information supplied for each reel. See Appendix B.